30

22S:30/105, Statistical Methods and Computing
Spring 2008, Instructor: Cowles
Midterm 3

Show your work on any problems that involve calculations.

Name: Solutions Course no. (30 or 105) _____

1. The U.S. Federal Highway Administration conducts annual surveys on motor vehicle travel by type of vehicle and publishes its findings in Highway Statistics. Independent random samples of cars, buses, and trucks provided the data on number of miles driven last year, in thousands. A few of the observations are shown below.

```
car 19.9
car 15.3
car  1.1
bus  1.8
bus  25.4
truck 24.6
truck 26.0
```

   (a) You wish to use this dataset to determine whether the population means of miles driven last year are the same for cars, busses, and trucks. Briefly explain why ANOVA, rather than the Chi square test, would be appropriate for your analysis.

3 ANOVA is appropriate for comparing means of a quantitative variable (miles driven) in 3 or more populations (cars, busses, trucks). X² is for proportions.

   (b) Refer to the SAS output attached. Does it provide any evidence that the assumptions required for the use of ANOVA are violated? Explain.

3 Yes. There are outliers in all 3 samples, the most severe in trucks. This suggests the population distributions are not normal. Sample standard deviations do not give evidence that pop. standard devs are not equal: largest $s = 13.07 < 2 \times$ smallest $= 2 \times 8.28$.

2. Hepatitis C causes about 10,000 deaths each year in the U.S. A person can be infected with hepatitis C for years before beginning to show any symptoms. A study from the University of Texas Southwestern Medical Center examined whether the risk of hepatitis C was related to whether people had tattoos and to where they got their tattoos. The 626 study participants were people being treated for a disease unrelated to hepatitis C. Although they were not a simple random sample, the way in which they were selected was unlikely to introduce severe bias into a study of hepatitis C. They were cross-classified in the following table:

6

	Hepatitis C	No Hepatitis C	Total
Tattoo, parlor	17	35	52
Tattoo, elsewhere	8	53	61
No tattoo	22	491	513
Total	47	579	626

(a) The researchers were interested in determining whether the proportion of people with Hepatitis C is the same in the three populations consisting of all people who get tattoos in tattoo parlors, all people who get tattoos elsewhere, and all people who have no tattoos. Why is the Chi square test, rather than ANOVA, appropriate for an analysis of this kind?

3

χ^2 test is appropriate when the variable of interest is binary (Hep C – yes/no) and we are interested in comparing proportions in 2 or more populations.

(b) List the rules of thumb for determining whether inference from the Chi square test are trustworthy. For each one, state whether this dataset satisfies the rule of them. Refer to the attached SAS output if it helps.

5

Population size > 10 × sample size in all samples. Yes, there are more than 520, 610 and 5130 people in the respective populations.

SRS's – no way to tell from info given.

Expected counts ≥ 5 in all cells. No! Two cells have expected counts < 5.

3. A marketing researcher was interested in people's recall of the brand names of products advertised on television. He assigned the participants in his study to watch a particular television program, during which 9 advertisements appeared. After the program ended, each subject was asked to write down the brand names of all the products that had been advertised. Twenty-four hours later, the people were again asked to write down the brand names of all the products that had been advertised. He wishes to use his data to determine whether the mean number of brands people recall is different immedately after the program versus 24 hours later. The sampling design in this study is (circle one):

2

(a) one sample
(b) two independent sample
(c) paired sample
(d) none of the above

4. An environmental scientist was interested in whether the proportion of cities that have recycling programs is different in the South from in the Midwest. She took a

10

random sample of cities from each of the two geographic regions. She found that 6 out of 16 southern cities had recycling programs, and 8 out of 14 midwestern cities had recycling programs.

(a) Write the scientist's null and alternative hypotheses. Use the standard symbols that we have used in class.

2 $H_0: p_s = p_{mw}$

$H_A: p_s \neq p_{mw}$

(b) Calculate a 90% confidence interval for the difference between the population proportions. Use the "plus 4" method. (Numeric answer; show your work.)

3 $\tilde{p}_s = \frac{7}{18} = 0.389$ $\tilde{p}_{mw} = \frac{9}{16} = 0.563$

$0.389 - 0.563 \pm 1.645\sqrt{\frac{0.389(0.611)}{18} + \frac{0.563(0.437)}{16}}$

$(-0.452, 0.104)$

(c) Use the result from the previous question to determine whether the scientist should reject her null hypothes at the 0.1 significance level. Briefly explain. (If you weren't able to do the previous problem, pretend that the answer is (-0.16,0.08)).

2 We cannot reject the null hypothesis that the two proportions are equal, because the confidence interval for their difference includes 0.

(d) Set up the calculation of the z test statistic for the hypothesis test. Fill in all the right numbers in the right places in the formula. You do not have to actually compute the final numeric result.

3 pooled $\hat{p} = \frac{6+8}{16+14} = 0.467$

$$z = \frac{\frac{6}{16} - \frac{8}{14} - 0}{\sqrt{0.467(0.533)\left(\frac{1}{16} + \frac{1}{14}\right)}}$$

5. A researcher wishes to study whether the mean grade point average is the same among UI math majors, UI sociology major, UI music majors, and UI chemistry majors. He takes simple random samples from each of the four populations of interest, and determines the grade point average of each student in the sample. The researcher has checked his data and found no outliers or other problems.

(a) Write the null and alternative hypotheses that are of interest to the researcher. Use conventional symbols.

$H_0: \mu_{math} = \mu_{soc} = \mu_{music} = \mu_{chem}$

2 $H_A: \mu_{math} \neq \mu_{soc}$ and/or

12

(b) How many different *pairs* of means are there in this problem? (Numeric answer)

6

(c) The researcher's first step in analysis should be (circle one):

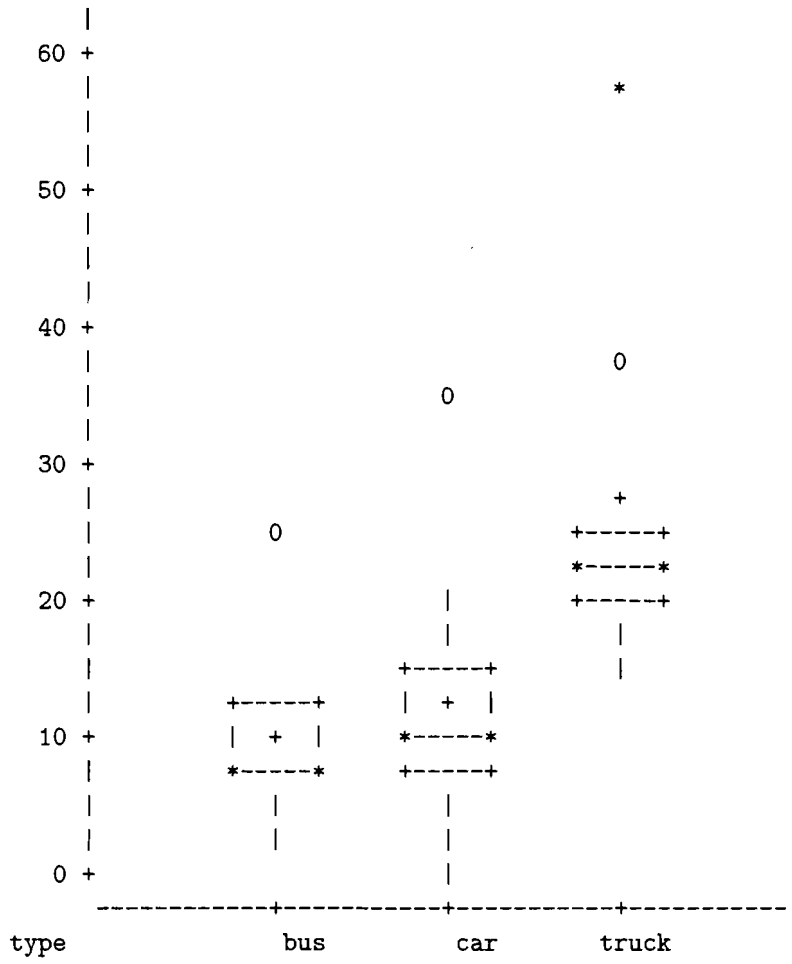
i. two-sample t-tests to determine in which pairs of majors the means are different

ii. an F test to determine whether there is evidence that any population means are different from each other

iii. a Chi square test to determine whether there is evidence that any population means are different from each other

iv. a z test

The UNIVARIATE Procedure
Variable: miles
Schematic Plots



Analysis Variable : miles

type	N Obs	N	Mean	Std Dev
bus	6	6	10.2333333	8.2812237
car	10	10	11.6400000	9.7219340
truck	9	9	26.9222222	13.0721821

Table of tattoo by hepC

tattoo	hepC		
Frequency / Expected / Percent / Row Pct / Col Pct	N	Y	Total
Else	53	8	61
	56.576	4.4244	
	8.18	1.23	9.41
	86.89	13.11	
	8.82	17.02	
None	513	22	535
	496.2	38.804	
	79.17	3.40	82.56
	95.89	4.11	
	85.36	46.81	
Parlor	35	17	52
	48.228	3.7716	
	5.40	2.62	8.02
	67.31	32.69	
	5.82	36.17	
Total	601	47	648
	92.75	7.25	100.00

Statistics for Table of tattoo by hepC

Statistic	DF	Value	Prob
Chi-Square	2	60.9869	<.0001
Likelihood Ratio Chi-Square	2	40.5105	<.0001
Mantel-Haenszel Chi-Square	1	12.2522	0.0005
Phi Coefficient		0.3068	
Contingency Coefficient		0.2933	
Cramer's V		0.3068	